Wine Quality Classification

CSC44700 - Final Project

Alice Liu

01

PROBLEM

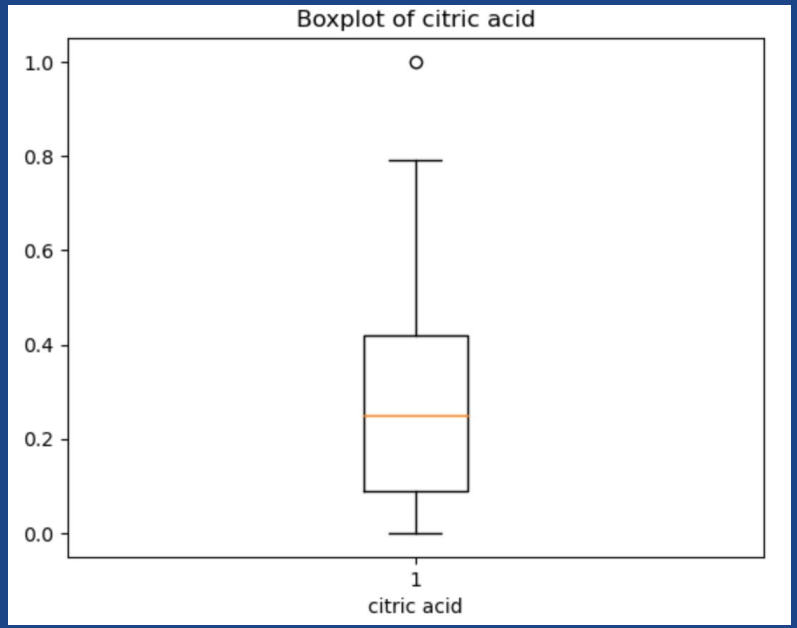
Build a model to predict the quality of wine based on various chemical attributes of the wine

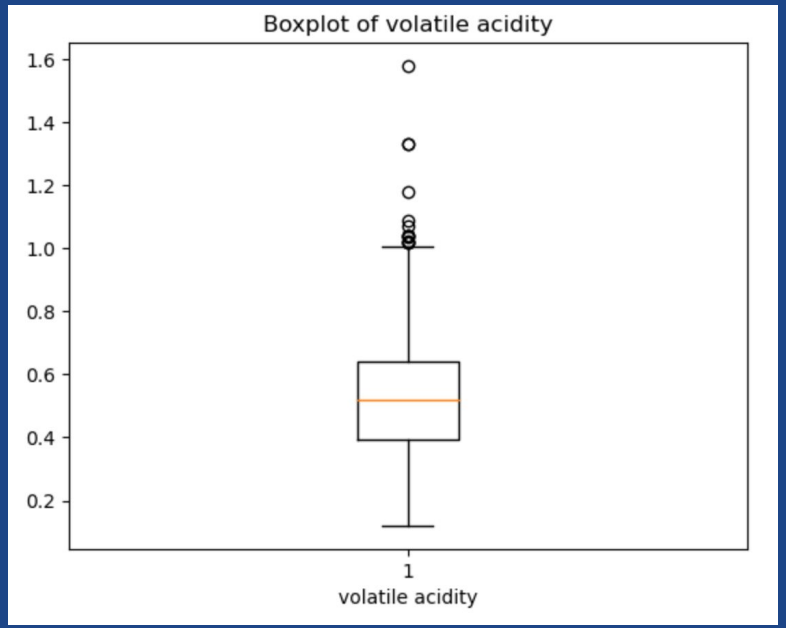
DATASET

Original Dataset

- Shape: (1143, 13)
- Columns: 13
  - ‘fixed acidity’, ‘volatile acidity’, ‘citric acidity’, ‘residual sugar’, ‘chlorides’, ‘free sulfur dioxide’, ‘total sulfur dioxide’, ‘density’, ‘pH’, ‘sulphates’, ‘alcohol’, ‘Id’
  - Target: ‘quality’
- No NULL

Initial EDA and Cleaning

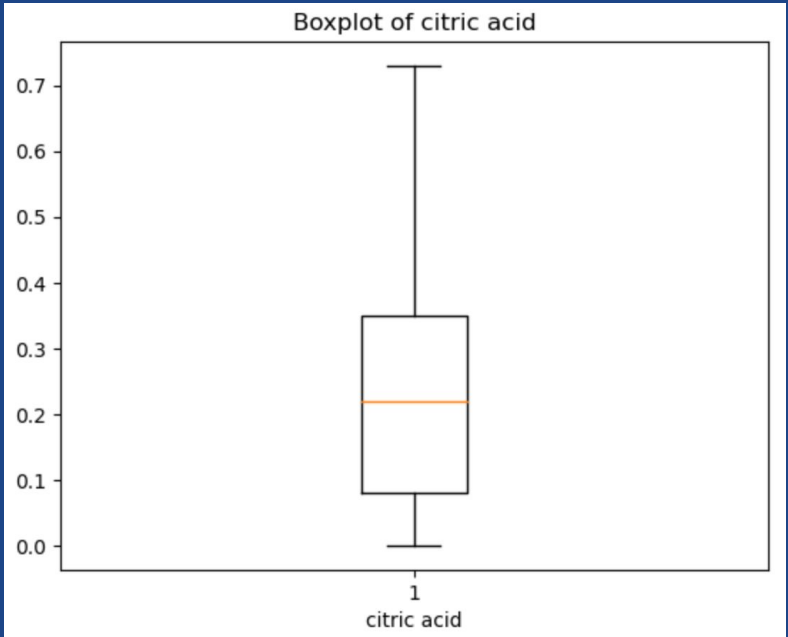


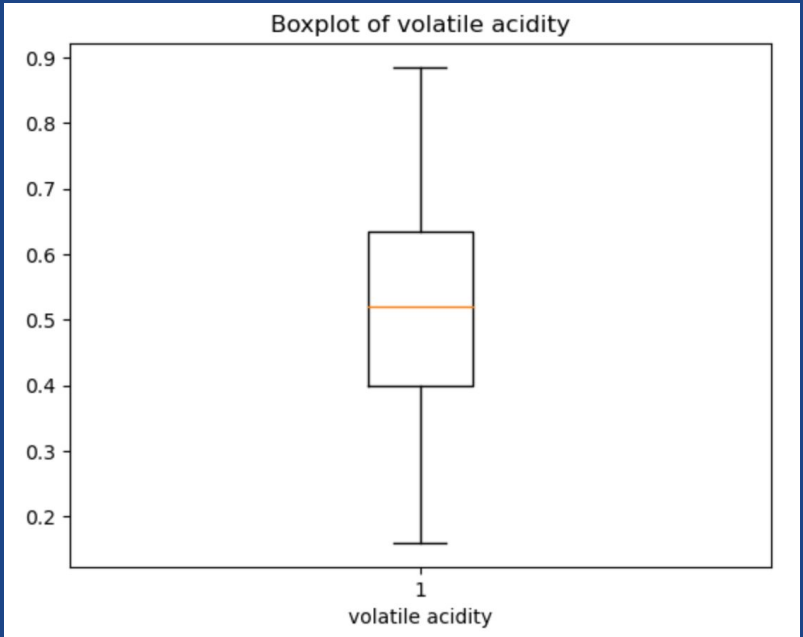


Initial EDA and Cleaning

```python
# DROP Outliers for all columns using IQR
def drop_outliers(df):
    df_new = pd.DataFrame()

    for col in df.columns:
        q1 = df[col].quantile(0.25)
        q3 = df[col].quantile(0.75)
        IQR = q3 - q1
        lower_bound = q1 - 1 * IQR  # 1.5 and 2 didn't work; 1 worked the best for threshold
        upper_bound = q3 + 1 * IQR  # 1.5 and 2 didn't work; 1 worked the best for threshold

        # drop outliers
        df_new[col] = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)][col]

    return df_new

df_clean = drop_outliers(df)
df_clean = df_clean.dropna()
plt_boxplots(df_clean)
```

Initial EDA and Cleaning



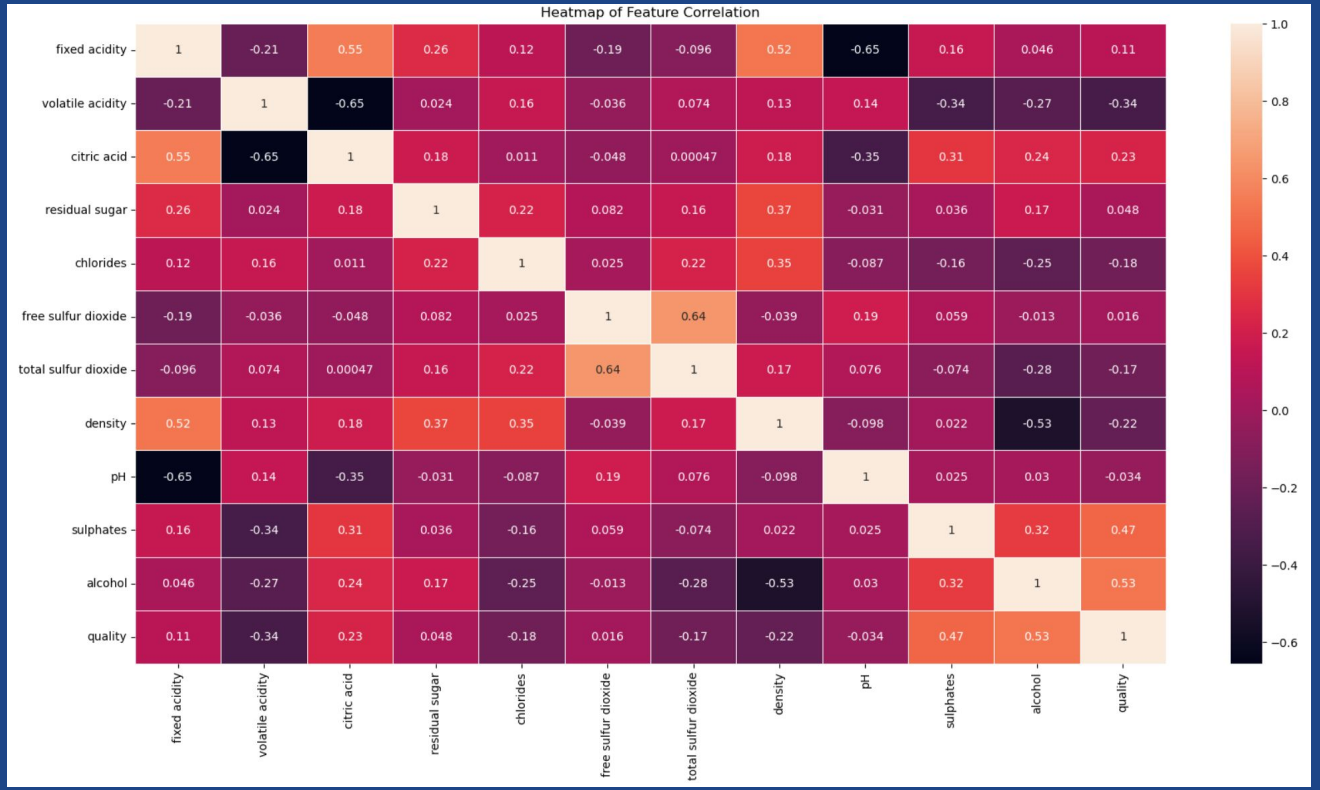


DATASET

Cleaned Dataset

- Shape: (633, 12)
- Columns: 12
  - ‘fixed acidity’, ‘volatile acidity’, ‘citric acidity’, ‘residual sugar’, ‘chlorides’, ‘free sulfur dioxide’, ‘total sulfur dioxide’, ‘density’, ‘pH’, ‘sulphates’, ‘alcohol’
  - Target: ‘quality’

More EDA



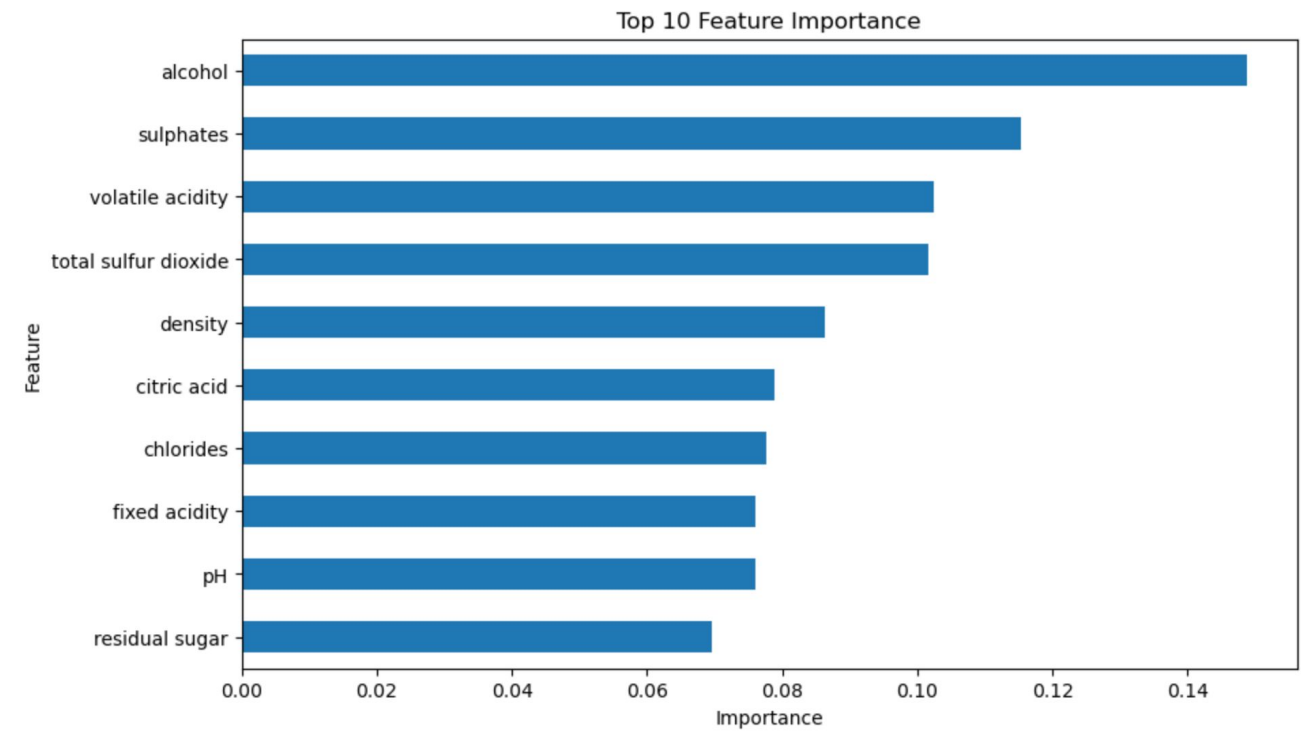
+ Correlation

- Alcohol vs Quality
- Sulphates vs Quality
- Citric Acid vs Fixed Acidity
- Density vs Fixed Acidity
- Free SO2 vs Total SO2

- Correlation

- Fixed Acidity vs pH
- Citric Acid vs Volatile Acidity
- Alcohol vs Density

FEATURE IMPORTANCE



MODELING

Logistic Regression

1. max_iter = 1000,
   random_state = 42
2. random_state = 42

Decision Trees

1. random_state = 42
2. criterion = gini,
   random_state = 42

Random Forest

1. n_estimators = 200,
   random_state = 42
2. random_state = 42

Gradient Boosting

1. n_estimators = 200,
   learning_rate = 0.1,
   random_state = 42

KNN

1. n_neighbors = 5
2. n_neighbors = 10

Extra Trees

1. n_estimators = 200

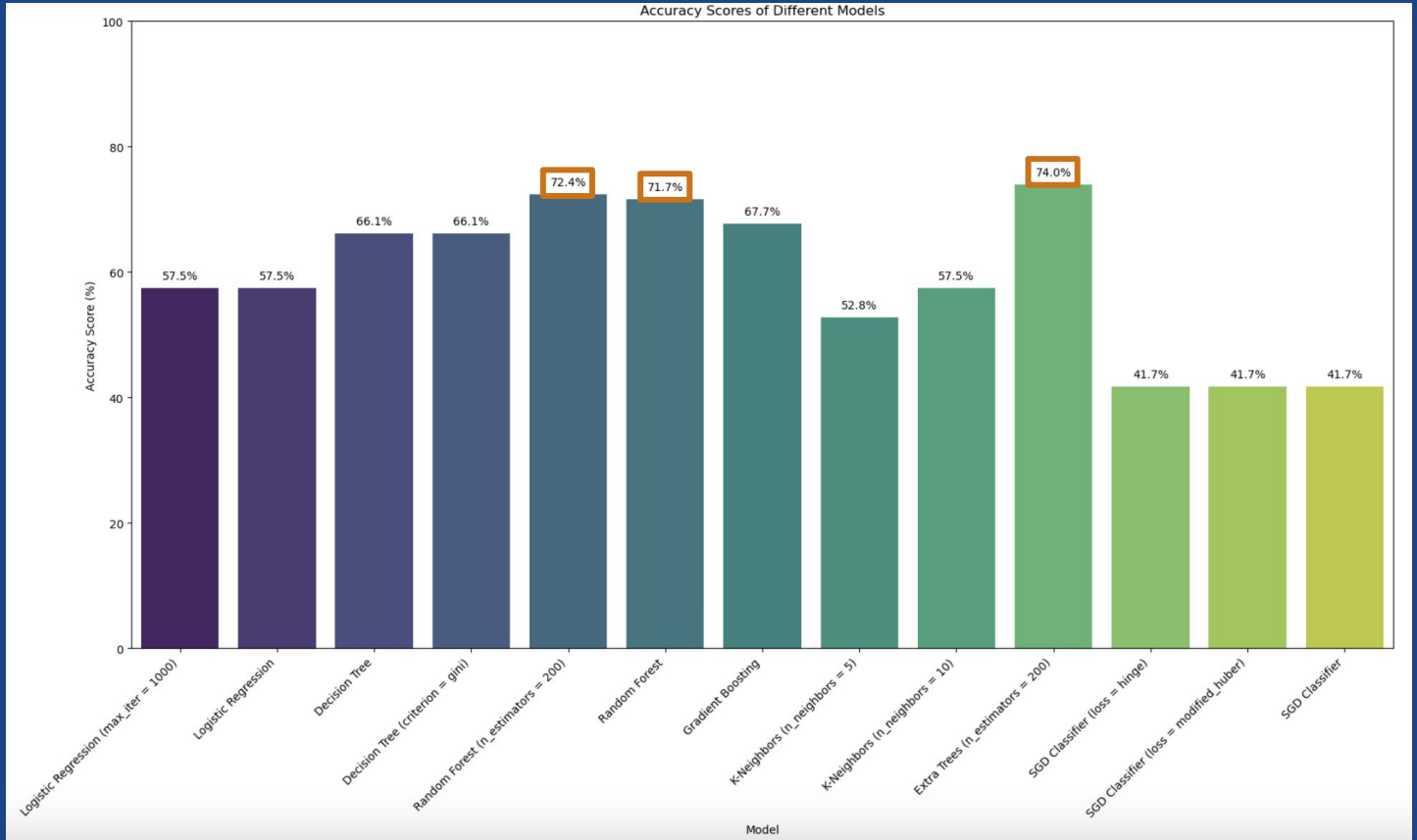
SGD Classifier

1. loss = 'hinge',
   random_state = 42
2. loss = 'modified_huber',
   random_state = 42

MODELING - Feature Selection

In [67]:
```
# using TOP 3 FEATURES FROM FEATURE IMPORTANCE <- 72% in RF, 74% in ExtraTrees
X_clean = df_clean[['alcohol', 'sulphates', 'volatile acidity']]
y_clean = df_clean['quality']

X_train_clean, X_test_clean, y_train_clean, y_test_clean = train_test_split(X_clean, y_clean,
                                                                            test_size = 0.2, random_state = 42)
```

EVALUATION

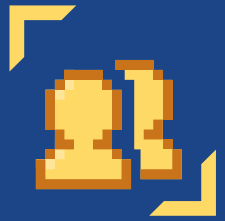

Accuracy Scores of Different Models
Accuracy Score (%)
Model
0
20
40
60
80
100
57.5%
57.5%
66.1%
66.1%
72.4%
71.7%
67.7%
52.8%
57.5%
74.0%
41.7%
41.7%
41.7%
Logistic Regression (max_iter = 1000)
Logistic Regression
Decision Tree
Decision Tree (criterion = gini)
Random Forest (n_estimators = 200)
Random Forest
Gradient Boosting
K-Neighbors (n_neighbors = 5)
K-Neighbors (n_neighbors = 10)
Extra Trees (n_estimators = 200)
SGD Classifier (loss = hinge)
SGD Classifier (loss = modified_huber)
SGD Classifier

HIGHEST: EXTRA TREES CLASSIFIER

74%

Accuracy Score

POINTS OF IMPROVEMENTS

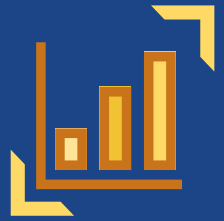


Feature Engineering

Explore additional feature engineering techniques (e.g., scaling, more effective handling of outliers, etc.)

Hyperparameter Tuning

Adjust the parameters, experiment using GridSearchCV



Model Selection

Try different models to see if they offer better performance, consider advanced ensemble techniques (e.g., stacking, blending multiple models)

THANK YOU!

Project GitHub Link:
https://github.com/AliceLiu17/Wine-Quality-Prediction